Modelling Referential Choice in Discourse: A Cognitive Calculative Approach and a Neural Network Approach[*]

André Grüning[1] and Andrej A. Kibrik[2]

[1]*Max Planck Institute for Mathematics in Sciences*
[2]*Institute of Linguistics, Russian Academy of Sciences*

In this paper, we discuss referential choice – the process of referential device selection made by the speaker in the course of discourse production. We aim at explaining the actual referential choices attested in the discourse sample. Two alternative models of referential choice are discussed. The first approach of Kibrik (1996, 1999, 2000) is the **cognitive calculative** approach. It suggests that referential choice depends on the referent's current activation score in the speaker's working memory. The activation score can be calculated as a sum of numeric contributions of individual activation factors, such as distance to the antecedent, protagonisthood, and the like. Thus, a predictive dependency between the activation factors and referential choice is proposed in this approach. This approach is cognitively motivated and allows one to offer generalization about the cognitive system of working memory. The calculative approach, however, cannot address non-linear interdependencies between different factors. For this reason we developed a mathematically more sophisticated **neural network** approach to the same set of data. We trained feed-forward networks on the data. They classified up to all but 4 instances correctly with respect to the actual referential choice. A pruning procedure allowed to produce a minimal network and revealed that out of ten input factors five were sufficient to predict the data almost correctly, and that the logical structure of the remaining factors can be simplified. This is a pilot study necessary for the preparation of a larger neural network-based study.

1 Introduction

We approach the phenomena of discourse reference as a realization of the process of *referential choice*: every time the speaker needs to mention a referent s/he has a variety of options at his/her disposal, such as full NPs, demonstratives, third person pronouns, etc. The speaker *chooses* one of these options according to certain rules that are a part of the language production system. Production-oriented accounts of reference are rarer in the literature than comprehension-oriented; for some examples see (Dale, 1992; Strube & Wolters, 2000).

Linguistic studies of referential choice often suffer from circularity: for example, a pronominal usage is explained by the referent's high activation,

---

while the referent is assumed to be highly activated because it is actually coded by a pronoun in discourse. In a series of studies by Kibrik (1996, 1999, 2000), an attempt to break such circularity was undertaken. The main methodological idea is that we need an account of referent activation that is entirely independent of the actual referential choices observed in actual discourse. There are a variety of linguistic factors that determine a referent's current activation, and once the level of activation is determined, the referential option(s) can be predicted with a high degree of certainty. This approach includes a quantitative component that models the interaction of activation factors yielding the summary activation of a referent. As it will be explained below, the contributions of individual factors are simply summed, and for this reason we use the shorthand *cognitive calculative approach*. This approach is outlined in Section 2 of this paper.

The cognitive calculative approach, however, has some shortcomings; in particular, its arithmetic nature could not allow addressing non-linear interaction between different factors. It is for this reason that we propose an alternative approach based on the mathematical apparatus of *neural networks*. In Section 3, computer simulations are reported in which we attempt to find out whether neural networks can help us to overcome some shortcomings of Kibrik's original approach. As the available data set is quite small (102 items) and large annotated corpora are not so easily obtained, we decided to design this study as a pilot study, rather than putting weight on statistical rigor.

2   The cognitive calculative approach

2.1   General assumptions underlying the cognitive calculative approach
In this paper, we approach discourse anaphora from the perspective of a broader process that we term referential device selection or, more simply, *referential choice*. This term differs from "discourse anaphora" in the following respects.

1) The notion of "referential choice" emphasizes the dynamic, procedural nature of reference in discourse. In addition, it is overtly production oriented: referential choice is the process performed by the speaker/writer. In the course of each act of referential choice, the speaker chooses a formal device to code the referent s/he has in mind. In contrast, "anaphora" is usually understood as a more static textual phenomenon, as a relationship between two or more segments of text.

2) Unlike "discourse anaphora", "referential choice" does not exclude introductory mentions of referents and other mentions that are not based on already-high activation of the referent.

3) The notion of referential choice permits one to avoid the dispute on whether "anaphora" is restricted to specialized formal devices (such as pronouns) or has a purely functional definition.

These three considerations explain our preference for the notion of referential choice. Otherwise the two notions are fairly close in their denotation.

A number of general requirements towards the cognitive calculative approach to referential choice were adopted from the outset of the study. The model must be:

(i)   speaker-oriented: referential choice is viewed as a part of language production performed by the speaker;
(ii)   sample-based: the data for the study is a sample of natural discourse, rather than heterogeneous examples from different sources;
(iii)   general: all occurrences of referential devices in sample must be accounted for;
(iv)   closed: the proposed list of factors cannot be supplemented to account for exceptions;
(v)   predictive: the proposed list of factors aims at predicting referential choice with maximally attainable certainty;
(vi)   explanatory and cognitively based: it is claimed that this approach models the actual cognitive processes, rather than relying on a black box ideology;
(vii)   multi-factorial: potential multiplicity of factors determining referential choice is recognized; each factor must be monitored in each case, rather than in an ad hoc manner, and the issue of interaction between various relevant factors must be addressed;
(viii) calculative: contributions of activation factors are numerically characterized;
(ix)   testable: all components of this approach are subject to verification;
(x)   non-circular: factors must be identified independently of the actual referential choice.

2.2   The cognitive model

Now, a set of more specific assumptions on how referential choice works at the cognitive level is in order. Recently a number of studies have appeared suggesting that referential choice is directly related to the more general cognitive domain of working memory and the process of *activation in working memory* (Chafe, 1994; Tomlin & Pu, 1991; Givón, 1995; Cornish, 1999; Kibrik, 1991, 1996, 1999). For cognitive psychological and neurophysiological accounts of working memory, see (Baddeley, 1986, 1990; Anderson, 1990; Cowan, 1995; Posner & Raichle, 1994; Smith & Jonides, 1997). The claim that referential choice is governed by memory processes is compatible with psycholinguistic frameworks of such authors as Gernsbacher (1990), Clifton and Ferreira (1987), Vonk *et al.* (1992), with the cognitively-oriented approaches of the Topic continuity research (Givón, 1983), Accessibility theory (Ariel, 1990), Centering theory (Gordon *et al.*, 1993), Givenness hierarchy

(Gundel *et al.*, 1993), and Cognitive grammar (van Hoek, 1997), as well as with some computational models covered in (Botley & McEnery, 2000). Thus, the first element of the cognitive model can be formulated as follows:

➢ The primary cognitive determiner of referential choice is activation of the referent in question in the speaker's working memory (henceforth: WM).

Activation is a matter of degree. Some chunks of information are more central in WM while some others are more peripheral. The term *activation score* (AS) is used here to refer to the current referent's level of centrality in the working memory. AS can vary within a certain range – from a minimal to a maximal value. This range is not continuous in the sense that there are certain important thresholds in it. When the referent's current AS is high, semantically *reduced* referential devices, such as pronouns and zeroes, are used. On the other hand, when the AS is low, semantically *full* devices such as full NPs are used. Thus the second basic idea of the cognitive model proposed here is the following.

➢ If AS is above a certain *threshold*, then a semantically reduced (pronoun or zero) reference is possible, and if not, a full NP is used.

Thus at any given moment in discourse any given referent has a certain AS. The claim is that AS depends on a whole gamut of various factors that can essentially be grouped in two main classes:

• *properties of the referent* (such as the referent's animacy and centrality)
• *properties of the previous discourse* (distance to the antecedent, the antecedent's syntactic and semantic status, paragraph boundaries, etc.)

These factors are specified below in Sections 2.3 and 2.4. Now the third basic point of the model can be formulated:

➢ At any given point of discourse *all relevant factors interact* with each other, and give rise to the integral characterization of the given referent (AS) with respect to its current position in the speaker's WM.

In other words, such oft-cited factors of referential choice as distance to the antecedent, referent centrality, etc., affect the referential choice not directly but through the mediation of the speaker's cognitive system, specifically, his/her WM. Therefore, these factors can be called *activation factors*.

The actual cognitive on-line process of referential choice is a bit more complex than is suggested by the three postulates formulated above. Some work on referential choice (see e.g. (Kibrik, 1991)) has been devoted to the issue of ambiguity of reduced referential devices. In the process of referential choice, a normal speaker filters out those referential options that can create ambiguity, or *referential conflict*. Thus, it is possible that even in case of high activation of a referent, a reduced referential device is still ruled out. The

referential conflict filter is outside of the focus of this paper, but consider one illustrative example from the Russian story discussed in the following section, in an English translation.

(1)   The mechanic started, but immediately returned – *he* began to dig in the box of
        instruments; they were lying in their places, in full order. *He* pulled out one wrench,
        dropped it, shook *his* head, whispered something and reached in again. Fedorchuk now
        clearly saw that the mechanic was a coward and would never go out to the wing. The
        pilot angrily poked **the mechanic** at the helmet with his fist <…>

The referent of interest here is "the mechanic"; all of its mentions are underlined, and the pronominal mentions are also italicised. The point in question is the boldfaced mention of this referent. "The mechanic" is very highly activated at this point (see Section 2.3 below), therefore, the pronominal mention ***him*** can be expected here. However, in the Russian original text (as well as in its English translation) such pronominal mention does not really fit. The reason is that, in spite of the extremely high activation of the referent, there is also at least one other referent, "Fedorchuk", that is equally activated and therefore can be assumed by the addressee to be the referent of the pronoun. Using a pronoun to refer to "the mechanic" would cause a referential conflict. Normally speakers/writers filter out the instances of potential referential conflict, by using disambiguation devices – from gender-specific pronouns to full NPs, as in example (1) (for details see (Kibrik, 1991, 2001)).
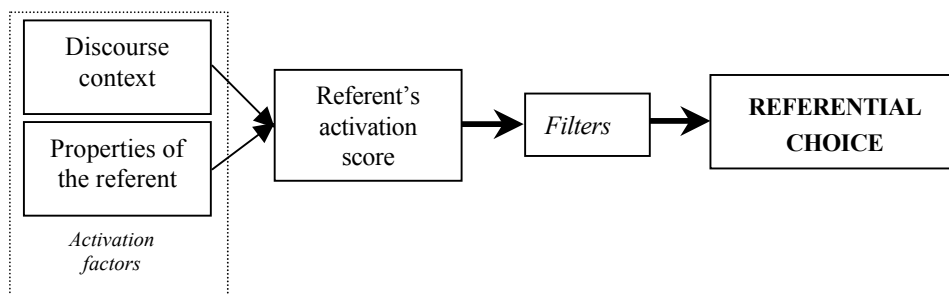

Figure 1: The cognitive multifactorial model of reference in discourse production

The cognitive model outlined above is summarized in the chart in Figure 1. The "filters" component implies, in the first place, the referential conflict filter, as well as some other filters, see (Kibrik, 1999).

This cognitive model is proposed here not only in a declarative way; there is also a mathematical, or at least quantitative, or *calculative* component to it. Each activation factor is postulated to have a certain numeric weight that reflects its relative contribution to the integral AS value. The general model of referential choice outlined above is assumed to be universal but the set of

activation factors, especially their relative numeric weights, and thresholds in the AS range are language-specific. In this article, two studies are reported that have been conducted for Russian (Section 2.3) and English (Section 2.4) written narrative discourse, with the explanation of the quantitative component of the model.

Both of the presented studies are based on small datasets, especially by standards of modern computational and corpus linguistics. However, it must be made clear that the original purport of these studies was of theoretical, rather than of computational, character: to overcome two major stumbling blocks common for the studies of reference. To reiterate, these two stumbling blocks are:

- *circularity*: Referential choice is explained by the level of activation (or another quasi-synonymous status), and the judgment on the level of activation is obtained from the actual referential form employed
- *multiplicity of factors*: Suppose factor A is of central importance in instance X, and factor B in instance Y. It often remains unclear what, if any, is the role of factor A in instance Y, and of factor B in instance X.

So, the goal of the proposed approach is to explore the following issue: is it possible to construct a system of activation factors that, first, are determined independently of actual referential choice, and, second, predict and explain referential choice in a cognitively plausible way?

As it will become clear from the exposition of the calculative component of this approach, this component is extremely time- and effort-consuming, and inherently must have been restricted to a small dataset. We believe this does not call into question the theoretical result: a system of interacting activation factors can indeed be constructed.

2.3   The Russian study

In this study (for details, see (Kibrik, 1996)), a single sample of narrative prose was investigated – a short story by the Russian writer Boris Zhitkov "Nad vodoj" ("Over the water"). This particular sample discourse was selected for this study because narrative prose is one of the most basic discourse types,[1] because written prose is a well-controlled mode in the sense that previous discourse is the only source for the recurring referents, and because Boris

---

[1] There is an unresolved debate in reference studies on whether referential processes are genre-dependent. Fox (1987a) proposed two different systems of referential choice, depending on discourse type. Toole (1996) has argued that the factors of referential choice are genre-independent. We do not address this issue in this article, but assume that in any case referential choice in narrative discourse must be close to the very nuclear patterns of reference, since narration is among the basic functions of language, is attested universally in all languages and cultures, and provides a very favorable environment for recurrent mention of referents in successive discourse units.

Zhitkov is an excellent master of style, with a very simple and clear language, well-motivated lexical choices, and at the same time with a neutral, non-exotic way of writing. This specific story is a prototypical narrative describing primarily basic events – physical events, interactions of people, people's reflections, sentiments, and speech. The story is written in the third person, so there are no numerous references to the narrator.

The sample discourse comprised about 300 discourse units (roughly, clauses). There are about 500 mentions of various referents in the sample, and there are some 70 different referents appearing in the discourse. However, only a minority of them occurs more than once. There are 25 referents appearing at least once in an anaphoric context, that is in a situation where at least a certain degree of activation can be expected.

The fundamental opposition in Russian referential choice is between full NPs and the third person pronoun *on*. Discourse-conditioned referential zeroes are also important, but they are rarer than *on* (for further details, see (Kibrik, 1996)).

Several textual factors have been suggested in the literature as directly determining the choice of referential device. Givón (1983, 1990) suggested that *linear distance* from an anaphor to the antecedent is at least one of the major predictors of referential choice. Givón measured linear distance in terms of clauses, and that principle turned out to be very productive and viable. In many later studies, including this one, discourse microstructure is viewed as a network of discourse units essentially coinciding with clauses. (There are certain reservations regarding this coincidence, but they are irrelevant for this paper.)

Fox (1987a: Ch. 5) argued that it is the rhetorical, hierarchical structure of discourse rather than plain linear structure that affects selection of referential devices. Fox counted *rhetorical distance* to the antecedent on the basis of a rhetorical structure constructed for a text in accordance with the Rhetorical Structure Theory (RST), as developed by Mann and Thompson (Mann *et al.*, 1992). According to RST, each discourse unit (normally a clause) is connected to at least one other discourse unit by means of a rhetorical relation, and via it, ultimately, to any other discourse unit. There exists a limited (although extensible) inventory of rhetorical relations, such as joint, sequence, cause, elaboration, etc. In terms of RST, each text can be represented as a tree graph consisting of nodes (discourse units) and connections (rhetorical relations). Rhetorical distance between nodes A and B is then the number of horizontal steps one needs to make to reach A from B along the graph (one example of a rhetorical graph is shown below in Section 2.4). Fox was correct in suggesting

that rhetorical distance measurement is a much more powerful tool for modelling reference than linear distance. However, linear distance also plays its role, though a more modest one.

In a number of works, it was suggested that a crucial factor of referential choice is episodic structure, especially in narratives. Marslen-Wilson *et al.* (1982), Tomlin (1987), and Fox (1987b) have all demonstrated, though using very different methodologies, that an episode/paragraph boundary is a borderline after which speakers tend to use full NPs even if the referent was recently mentioned. Thus, one can posit the third type of distance measurement – *paragraph distance*, measured as the number of paragraph boundaries between the point in question and the antecedent.

One more factor was emphasized in (Grimes, 1978) – the centrality of a referent in discourse, which we call *protagonisthood* below. For a discussion of how to measure a referent's centrality, see (Givón, 1990: 907-909).

Several other factors have been suggested in the literature, including animacy, syntactic and semantic roles played by the NP/referent and by the antecedent, distance to the antecedent measured in full sentences, and the referential status of the antecedent (full/reduced NP). Some of these factors will be discussed in greater detail in Section 2.4 below, in connection with the English data.

From the maximal list of potentially significant activation factors we picked a subset of those that proved actually *significant* for Russian narrative prose. The criterion used is as follows. Each factor can be realized in a number of values. For example, a distance factor may have values 1, 2, etc. Each potentially significant factor has a "privileged" value that presumably correlates with the more reduced form of reference. For example, for the linear distance to the antecedent, it is the value of "1", while for the factor of the antecedent's syntactic role it is "subject". Only those potential factors whose privileged value demonstrated a high co-occurrence (in at least 2/3 of all cases) with the reduced form of reference have been considered significant activation factors. For example, the factor of rhetorical distance patterns vis-à-vis pronouns and full NPs in a nearly mirror image way: there is a high co-occurrence of the value of 1 with pronominal reference (91%), and a high co-occurrence of rhetorical distance greater than 1 (79%) with full NP reference.

On the other hand, other potential factors did not display any significant co-occurrence with referential choice. In particular, the parameter of referential type of the antecedent does not correlate at all with the referent's current pronominalizability: for instance, a 3rd person pronoun is the antecedent of 10%

of all 3rd person pronouns and 13% of full NPs, which makes no significance difference.

Seven significant activation factors have been detected. This is their list with the indication [in brackets] of the privileged value co-occurring with pronominal reference: animacy [human], protagonisthood [yes], linear distance [1], rhetorical distance [1], paragraph distance [0], syntactic [subject] and semantic [Actor[2]] roles of the antecedent, and sloppy identity[3].

After the set of significant activation factors had been identified, certain numeric weights have been assigned to their values. Variation of referents' AS from 0 to 1 was postulated. The activation factor weights take discrete values measured in steps of size 0.1. In each particular case, the weights of all involved factors can be summed and the resulting activation score is supposed to predict referential choice.

Table 1 below lists a selection of activation factors. Each factor is presented with the values it can accept and the corresponding numeric weights.

Activation factor	Value	Numeric activation weight
Rhetorical distance to the antecedent	1	0.7
	2	0.4
	3	0
	4+	−0.3
Paragraph distance to the antecedent	0	0
	1	−0.2
	2+	−0.4
Protagonisthood	Yes, and the current mention is:	
	the 1st mention in a series	0.3
	the 2nd mention in a series	0.1
	otherwise	0
	No	0

Table 1: Examples of activation factors, their values, and numeric weights

Activation factors differ regarding their logical structure. Some factors are sources of activation. The strongest among these is the factor of rhetorical

---

[2] The term "Actor" is an abstract semantic macrorole; it designates the semantically central participant of a clause, with more-than-one-place verbs usually agent or experiencer; see e.g. (Van Valin, 1993:43ff).

[3] The factor of sloppy identity occurs when two expressions are referentially close, but not identical. In the following example from the story under investigation, given in a nearly literal English translation, the first expression is referentially specific, and the second (_it_) generic:

    (i) He understood that the engine skipped, that probably the carburetor had gotten clogged (through _it_ gas
        gets into an engine) <…>

Sloppy identity is relevant in far fewer cases than other factors, and for this reason it can be called a second-order, or "weak", factor. Sloppy identity slightly reduces activation of a referent that has an antecedent, but a sloppy one.

distance to the antecedent. The closer the rhetorical antecedent is, the higher is the activation.

The factor of paragraph distance is never a source of activation; vice versa, it is, so to speak, a penalizing factor. In the default situation, when the antecedent is in the same paragraph (paragraph distance = 0), this factor does not contribute to AS at all. When the antecedent is separated from the current point in discourse by one or more paragraph boundaries, the activation is lowered.

The third factor illustrated in Table 1, protagonisthood, has a still different logical structure. It can be called a compensating factor. It can only add activation, but does that in very special situations. When a referent is not a protagonist, this factor does not affect activation. If a referent is a protagonist, this factor helps to regain activation at the beginning of a series,[4] that is, in the situation of lowered activation. If the activation is high anyway, this factor does not matter.

The numeric weights such as those in Table 1 were obtained through a heuristic procedure of trials and errors. After several dozen of successive adjusting trials the numeric system turned out to predict a subset of referential choices correctly: reduced referential forms were getting ASs close to 1, and full NPs were getting ASs much closer to 0. When this was finally achieved, it turned out that all other occurrences of referential devices are properly predicted by this set of numeric weights without any further adjustment. It is worth pointing out that such trial-and-error procedure, performed by hand, is extremely time- and labour-consuming, even provided that the dataset was relatively small. The difference of this approach from the prior approaches is that the full control of the dataset, whatever size it has, has been gained.

After the calculative model was completely adjusted to the data of the Zhitkov's story, it was tested on a different narrative – a fragment of Fazil' Iskander's story "Stalin and Vuchetich", about 100 discourse units long. The result was that the model predicted all referential choices in the test dataset, without further adjustment (with the exception of minor adjustment in the numerical weights of two activation factors). These facts can be taken as evidence suggesting that the developed system does model actual referential choice in written narratives closely enough.

One more crucial point needs to be made about this model. When one observes actual referential choices in actual discourse, one can only see the ready results of referential device selection by the author – full NPs, pronouns,

---

[4] The notion of "series" means a sequence of consecutive discourse units, such that: (i) all of them mention the referent in question, and (ii) the sequence is preceded by at least three consecutive discourse units not mentioning the referent.

or zeroes. However, the real variety of devices is somewhat greater. It is important to distinguish between the *categorical* and *potentially alternating* referential choices. For example, the pronoun *on* in a certain context may be the only available option, while in another context it could well be replaced by an equally good referential option, say a full NP. These are two different classes of situations, and they correspond to two different levels of referent activation. The referential strategies formulated in (Kibrik, 1996) for Russian narrative discourse are based on this observation. Those referential strategies shown in Table 2 below represent the mapping of different AS levels onto possible referential choices.

Referential device:	Full NP only	Full NP most likely, pronoun /zero unlikely	Either full NP or pronoun/zero	Pronoun/zero only
AS:	0–0.3	0.4–0.6	0.7–0.9	1

Table 2: Referential strategies in Russian narrative discourse

What governs the speaker's referential choice when the AS is within the interval of the activation scale that allows variable referential devices (especially 0.7 through 0.9)? We do not have a definitive answer to this question at present. The choice may depend on idiolect, on discourse type and genre, or perhaps even be random. On the other hand, there may be some additional, extra-weak, factors that come into play in such situations.

2.4 The English study

The model developed for Russian narrative discourse was subsequently applied to a sample of English narrative discourse, which required a fair amount of modification. This study was described in (Kibrik, 1999), and its main results are reported here, along with some additional details. The sample (or small corpus) was the children's story "The Maggie B." by Irene Haas. There are 117 discourse units in it. 76 different *referents* are mentioned in it, not counting 13 other mentioned in the quoted songs. There are 225 referent *mentions* in the discourse (not counting those in quoted text). There are 14 different referents mentioned in discourse that are *important* for this study. They are those mentioned at least once in a context where any degree of activation can be possibly expected. Among the important referents, there are three protagonist referents: "Margaret" (72 mentions altogether), "James" (28 mentions), and "the ship" (12 mentions). An excerpt from the sample discourse, namely lines 1401–2104, is given in the Appendix below.

Any referent, including an important referent, can be mentioned in different ways, some of which (for example, first person pronouns in quoted speech) are irrelevant for this study. Those that are *relevant* for this study fall into two large formal classes: references by full NPs and references by activation-based

pronouns. "Activation-based pronouns" means the unmarked, general type of pronoun occurrences that cannot be accounted for by means of any kind of syntactic rules, in particular, for the simple reason that they often appear in a different sentence than their antecedents. In order to explain and predict this kind of pronoun occurrence, it is necessary to construct a system of the type described in Section 2.3, taking into account a variety of factors related to discourse context and referents' properties. Typical examples of activation-based pronouns are given in (2) below[5].

(2) 1607   Lightning split the sky
     1608   as _she_ ran into the cabin
     1609   and slammed the door against the wet wind.
     1610   Now everything was safe and secure.
     1701   When _she_ lit the lamps,
     1702   the cabin was bright and warm.

There are two occurrences of the activation-based pronoun _she_ in (2), and the second one is even used across the paragraph boundary from its antecedent. Besides the activation-based 3rd person pronouns, there are two dozen occurrences of syntactic pronouns that can potentially be accounted for in terms of simpler syntactic rules. At the same time, the activation-based principles outlined here can easily account for syntactic pronouns, see (Kibrik, 1999).[6]

Thus, the focus of this study was restricted to 39 full NP references and 40 activation-based pronominal references. As it was pointed out in Section 2.3 above, within each of the referential types – full NPs and pronouns – there is a crucial difference: whether the referential form in question has _an alternative_. In (3) below, an illustration of a pronoun usage is given that can vary with a full NP: in unit 1601, the full NP Margaret could well be used (especially provided that there is a paragraph boundary in front of unit 1601).

(3) 1502   A storm was coming!
     1503   Margaret must make the boat ready at once.
     1601   _She_ took in the sail
     1602   and tied it tight.

---

[5] In the examples, as well as in Appendix 1, each line represents one discourse unit. In line numbers the first two digits refer to the paragraph number in the story, and the last two digits to the number of the discourse unit within the current paragraph.

[6] For an example of a syntactic pronoun cf. one sentence from the story under investigation (see Appendix, lines 1601-1602):

   (ii) She took in the sail and tied _it_ tight.

Pronouns occurrences such as _it_ in this example can be accounted for by means of syntactic rules that are lighter, in some sense, than the activation-based procedure of referential choice described here. For an example of a generalized treatment of activation-based and syntactic referential devices see Section 3 of this article.

Contrariwise, there are instances of categorical pronouns. Consider (4), which is a direct continuation of (3):

(4) 1603   _She_ dropped the anchor
      1604   and stowed all the gear <...>

In 1603, it would be impossible to use the full NP Margaret; only a pronoun is appropriate.

For the English data, it was found that referential forms of each type (for example, pronouns) fall into three categories: those allowing no alternative (= categorical), those allowing a questionable alternative, and those allowing a clear alternative. Thus, there are six possible correspondences between the five potential types and two actual realizations; see Table 3.

Potential referential form	Full NP only	Full NP, $^?$pronoun	Full NP or pronoun	Pronoun, $^?$full NP	Pronoun only
Frequency	_15_	_17_	_7_	_15_	_18_	_7_

| Actual referential form | Full NP (_39_) | | | Pronoun (_40_) | |

Table 3: Actual and potential referential forms, and their frequencies in sample discourse

The information about referential alternatives is crucial for establishing referential strategies. Of course, attribution of particular cases to one of the categories is not straightforward. It must be noted that such attribution is the second extremely laborious procedure involved in this kind of study (along with the search for optimal numerical weights of activation factors). To do this attribution properly, a significant number of native speakers must be consulted. There were two sources of information on referential alternatives used in this study: (i) an expert who was a linguist and a native speaker of English and had a full understanding of the problem and the research method, and who supplied her intuitive judgments on all thinkable referential alternatives in all relevant points of discourse; (ii) a group of 12 students, native speakers of English, who judged the felicity of a wide variety of modifications of the original referential choices through a complicated experimental procedure. These two kinds of data were brought together and gave rise to an integral judgment for each referential alternative. The details of this part of the study are reported in (Kibrik, 1999). In the end, all referential alternatives were classified as either appropriate, questionable, or inappropriate – see Table 4 below. The attribution of referential alternatives to categories is an indispensable component of this study, since the two formal categories "pronoun" vs. "full NP" are far too rough to account for the actual fluidity of referential choice.

The six strongest activation factors that were found to be most important in modelling the data of the sample discourse are the following: rhetorical distance to the antecedent (RhD), linear distance to the antecedent (LinD), paragraph distance to the antecedent (ParaD), syntactic role of the linear antecedent,[7] animacy, and protagonisthood. The first three of these factors are different measurements of the distance from the point in question to the antecedent. By far, the most influential among the distance factors, and in fact among all activation factors, is the factor of *rhetorical distance*: it can add up to 0.7 to the activation score of a referent. *Linear* and *paragraph distances* can only penalize a referent for activation; this happens if the distance to the antecedent is too high. To see how rhetorical (hierarchical) structure of discourse can be distinct from its linear structure, consider the rhetorical graph in Figure 2.[8]

Rhetorical distance is counted as the number of horizontal steps required in order to reach the antecedent's discourse unit from the current discourse unit. For a simple example, consider the pronoun *him* in discourse unit 1802. It has its antecedent James in discourse unit 1801. There is one horizontal step from 1802 to the left to 1801, hence RhD = 1. The pronoun *they* in 2004 has its antecedent Margaret and James in 2001. In order to reach 2001 from 2004, one needs to make two horizontal steps along the tree leftwards: 2004 to 2002 and 2002 to 2001. To visualize this more clearly, it is useful to collapse the fragment of the tree onto one linear dimension, see Figure 3. Thus RhD = 2.

In narratives, the fundamental rhetorical relation is that of sequence. Three paragraphs of the four depicted in Figure 2 (#18, #20, and #21) are connected by this relation, and within each of these paragraphs there are sequenced discourse units, too. If there were no other rhetorical relations in narrative besides sequence, rhetorical distance would always equal linear distance. However, this is not the case. In the example analysed, one paragraph, namely #19, is off the main narrative line. It provides the background scene against which the mainline events take place. Likewise, discourse unit 1904 reports a result of what is reported in 1903. The difference between the linear and the rhetorical distance can best be shown by the example of discourse unit 2001. For the referents "Margaret" and "James", mentioned therein, the nearest antecedents are found in discourse unit 1802. It is easy to see that the linear distance from 2001 to 1802 is 6 (which is a very high distance) while the

---

[7] Note that one referent mention often has two distinct closest antecedents: a rhetorical and a linear one.

[8] It is a commonplace in the research on Rhetorical Structure Theory that there is certain constrained variation in how a given text can be represented as a hierarchical graph by different annotators (Mann *et al.*, 1992; Carlson *et al.*, 2003). To be sure, the fact of variation is the inherent property of discourse interpretation, and there is no other way of getting "better" hierarchical trees than rely on judgment of trained experts.

rhetorical distance is just 2 (first step: from 2001 to 1803, second step from 1803 to 1802). Perhaps the most conclusive examples of the power of rhetorical distance as a factor in referential choice are the cases of long quotations: it is often the case that in a clause following a long quotation one can use a pronoun, with the nearest antecedent occurring before the quotation. This is possible in spite of the very high linear distance, and due to the short rhetorical distance: the pronoun's clause and the antecedent's clause in such case can be directly connected in the rhetorical structure.
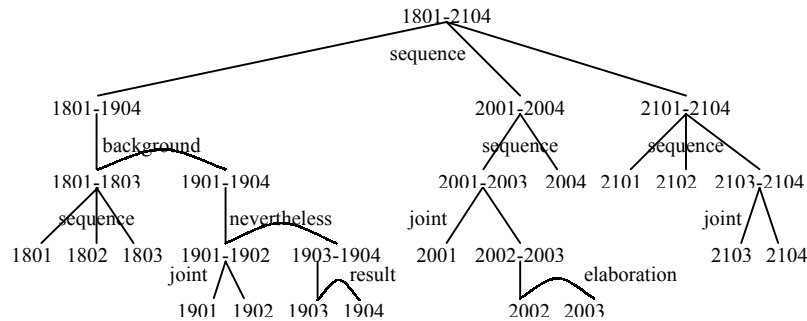
Figure 2: A rhetorical graph corresponding to lines 1801–2104 of the excerpt given in the Appendix
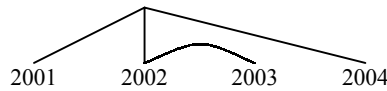
Figure 3. One-dimensional representation of a fragment of the rhetorical graph.

The following factor, indicated above, and the second most powerful source of activation, is the factor of *syntactic role of the linear antecedent*. This factor applies only when the linear distance is short enough: after about four discourse units it gets forgotten what the role of the antecedent was; only the fact of its presence may still be relevant. Also, this factor has a fairly diverse set of values. As it has long been known from studies of syntactic anaphora, subject is the best candidate for the pronoun's antecedent. (This observation is akin to the ranking of "forward-looking centres" in Centering theory, suggesting that the subject of the current utterance is the likeliest among other participants to recur in the next utterance with a privileged status; see e.g. (Walker & Prince, 1996: 297)). Different subtypes of subjects, however, make different contributions to referent activation, ranging from 0.4 to 0.2. Other relevant values of the factor include the direct object, the indirect (most frequently, agentive) object, the possessor, and the nominal part of the predicate. It is very typical of pronouns, especially for categorical pronouns (allowing no full NP alternative) to have subjects as their antecedents. For example, consider three pronouns in

paragraph #16 (see Appendix): *she* (discourse unit 1603), *her* (1606), and *she* (1608). According to the results of the experimental study mentioned above, the first and the second pronouns are categorical (that is, Margaret could not be used instead) and they have subject antecedents. But the third one has a non-subject antecedent, and it immediately becomes a potentially alternating pronoun (Margaret would be perfectly appropriate here).[9]

The following two factors are related not to the previous discourse but to the relatively stable properties of the referent in question. *Animacy* specifies the permanent characterization of the referent on the scale "human – animal – inanimate". *Protagonisthood* specifies whether the referent is the main character of the discourse. Protagonisthood and animacy are rate-of-deactivation compensating factors (see discussion in Section 2.3). They capture the observation that important discourse referents and human referents deactivate slower than those referents that are neither important nor human. In addition, a group of second-order, or "weak", factors were identified, including the following ones. *Supercontiguity* comes into play when the antecedent and the discourse point in question are in some way extraordinarily close (e.g. being contiguous words or being in one clause). *Temporal or spatial shift* is similar to paragraph boundary but is a weaker episodic boundary; for example, occurrence of the clause-initial *then* frequently implies that the moments of time reported in two consecutive clauses are distinct, in some way separated from each other rather than flowing one from the other. *Weak referents* are those that are not likely to be maintained, they are mentioned only occasionally. Such referents often appear without articles (cf. NPs *rain, cinnamon and honey, supper* in the text excerpt given in the Appendix) or are parts of stable collocations designating stereotypical activities (*slam the door, light the lamps, give a bath*). Finally, *introductory antecedent* means that when a referent is first introduced into discourse, it takes no less than two mentions to fully activate it.

For details on the specific values of all activation factors, and the corresponding numeric weights, refer to (Kibrik, 1999). As in case of the Russian study, the numeric activation weights of each value were obtained through a long heuristic trial-and-error procedure. All referential facts contained in the original discourse and obtained through experimentation with alternative forms of reference are indeed predicted/explained by the combination of activation factors with their numeric weights, and the referential strategies.

---

[9] This demonstration of one factor operating in isolation is not intended to be conclusive, since the essence of the present approach is the idea that all factors operate *in conjunction*. It does, however, serve to illustrate the point.

The referential strategies formulated in this study are represented in Table 4. As in Section 2.3, the referential strategies indicate the mappings of different intervals on the AS scale onto possible referential devices.

Referential device:	Full NP only	Full NP, $^?$pronoun	Either full NP or pronoun	Pronoun, $^?$full NP	Pronoun only
AS:	0–0.2	0.3–0.5	0.6–0.7	0.8–1.0	1.1+

Table 4: Referential strategies in English narrative discourse

The quantitative system in this study was designed so that AS can sometimes exceed 1 and reach the value of 1.1 or even 1.2. This is interpreted as "extremely high activation" (it gives the speaker no full NP option to mention the referent, see the value in the rightmost column of Table 4 and below). The AS of 1 is then interpreted as "normal maximal" activation. Also, a low AS frequently turns out to be negative. Such values are simply rounded to 0.

According to the referential strategies represented in Table 4, the five categories of potential referential forms correspond to five different intervals on the activation scale. There are four thresholds on this scale. The thresholds of 0.2 and 1.0 are hard: when the AS is 0.2 or less, a pronoun cannot be used, and when it is over 1.0, a full NP cannot be used. There are also two soft thresholds: when the AS is 0.5 or less, a pronoun is unlikely, and when it is over 0.7, a full NP is unlikely.

To demonstrate how predictively the calculative system of activation factors works, several examples of actual calculations are presented in Table 5. All examples are taken from the text excerpt given in the Appendix. Examples are different in that they pertain to different referential options possible on the AS scale (see Table 4 above). There is one example for each of the following referential options: (a) full NP, $^?$pronoun; (b) either full NP or pronoun; (c) pronoun, $^?$full NP; (d) pronoun only.

The upper portion of Table 5 contains a characterization of each example: its location in the text, the actual referential form used by the author, the referent, the type of referential device and possible alternative devices, as obtained through the experimental study described above. Also, the AS interval corresponding to the referential option in question is indicated, in accordance with the referential strategies given in Table 4 above. The lower middle portion of Table 5 demonstrates the full procedure of calculating the ASs, in accordance with the values' numeric weights. The last line of Table 5 indicates whether the calculated AS fits within the range predicted by the referential strategies.

Referential option	(a) Full NP, ?pronoun	(b) Full NP or pronoun	(c) Pronoun, ?full NP	(d) Pronoun only
Line number	1802	1701	1802	1603
Referential form	Margaret	*She*	*him*	*she*
Referent	"Margaret"	"Margaret"	"James"	"Margaret"
Actual referential device	full NP	pronoun	pronoun	pronoun
Alternative referential device	?pronoun	full NP	?full NP	—
Corresponding AS interval	*0.3–0.5*	*0.6–0.7*	*0.8–1.0*	*1+*
Relevant activation factors				
RhD VALUE:	3	2	1	1
NUM. WEIGHT:	*0*	*0.5*	*0.7*	*0.7*
LinD VALUE:	3	2	1	1
NUM. WEIGHT:	*–0.2*	*–0.1*	*0*	*0*
ParaD VALUE:	1	1	0	0
NUM. WEIGHT:	*–0.3*	*–0.3*	*0*	*0*
Lin. antec. role VALUE:	S	S	passive S	S
NUM. WEIGHT:	*0.4*	*0.4*	*0.2*	*0.4*
Animacy VALUE:	Human, $LinD \geq 3$	Human, $LinD \leq 2$	Human, $LinD \leq 2$	Human, $LinD \leq 2$
NUM. WEIGHT:	*0.2*	*0*	*0*	*0*
Protagonisthood VALUE:	Yes, $RhD+ParaD \geq 3$	Yes, $RhD+ParaD \geq 3$	Yes, $RhD+ParaD \leq 2$	Yes, $RhD+ParaD \leq 2$
NUM. WEIGHT:	*0.2*	*0.2*	*0*	*0*
Calculated AS	*0.3*	*0.7*	*0.9*	*1.1*
Fit within the predicted AS interval	*Yes*	*Yes*	*Yes*	*Yes*

Table 5: Examples of calculations of the referents' ASs in comparison with the predictions of the referential strategies (for explanation of factors' values see Kibrik 1999)

2.5 Consequences for working memory

The studies outlined in Sections 2.3 and 2.4 rely on work in cognitive psychology, but they are still purely linguistic studies aiming at explanation of phenomena observed in natural discourse. However, it turns out that the results of those studies are significant for a broader field of cognitive science, specifically for research in *working memory*.

Working memory (WM; otherwise called short-term memory or primary memory) is a small and quickly updated storage of information. The study of WM is one of the most active fields in modern cognitive psychology (for reviews see (Baddeley, 1986; Anderson, 1990: Ch. 6); some more recent approaches are represented in (Gathercole, 1996; Miyake & Shah, 1999; Schroeger *et al.*, 2000). WM is also becoming an important issue in neuroscience: see (Smith & Jonides, 1997). There are a number of classical issues in the study of WM. Shah and Miyake (1999) list eight of major

theoretical questions in WM. It appears that the results obtained in this linguistic study contribute or at least relate to the majority of these hot questions, including:

- *capacity*: how much information can there be in WM at one time?
- *forgetting*: what is the mechanism through which information quits WM?
- *control*: what is the mechanism through which information enters WM?
- *relatedness to attention*: how do WM and attention interact?
- *relatedness to general cognition*: how does WM participate in complex cognitive activities, such as language?
- *(non-)unitariness*: is WM a unitary mechanism or a complex of multiple subsystems?

Here, only some results related to the issues of capacity and attentional control will be mentioned. For more detail, refer to (Kibrik, 1999).

The system of activation factors and their numeric weights was developed in order to explain the observed and potential types of referent mentions in discourse. In the first place, only those referents that were actually mentioned in a given discourse unit were considered. But this system was discovered to have an additional advantage: it operates independently of whether a particular referent is actually mentioned at the present point in discourse. That is, the system can identify any referent's activation at any point in discourse no matter whether the author chose to mention it in that unit or not. If so, one can calculate the activation of *all referents* at a given point in discourse. Consider discourse unit 1608 (see Appendix). Only two referents are mentioned there: "Margaret" and "the cabin". However, the following other referents have AS greater than 0 at this point: "the anchor", "the gear", "rain", "the deck", "thunder", "lightning", and "the sky". The sum of ASs of all relevant referents gives rise to *grand activation* – the summed activation of all referents at the given point in discourse. Grand activation gives us an estimate of the *capacity* of the specific-referents portion of WM.

Figure 4 depicts the dynamics of activation processes in a portion of the English discourse (lines 1401 through 2104, see Appendix). There are three curves in Figure 4: two pertaining to the activation of the protagonists "Margaret" and "James", and the third representing the changes in grand activation. Observations of the data in Figure 4 make it possible to arrive at several important generalizations. Grand activation varies normally within the range between 1 and 3, only rarely going beyond this range and not exceeding 4. Thus the variation of grand activation is very moderate: maximally, it exceeds the maximal activation of an individual referent only about three to four times. This gives us an estimate of the maximal capacity of the portion of WM related to specific referents in discourse: three or four fully activated

referents. Interestingly, this estimate coincides with the results recently obtained in totally independent psychological research looking at working memories specialized for specific kinds of information (Velichkovsky *et al.*, 1995; Cowan, 2000). Furthermore, there are strong shifts of grand activation at paragraph boundaries; even a visual examination of the graph in Figure 4 demonstrates that grand activation values at the beginnings of all paragraphs are local minima; almost all of them are below 2. On the other hand, in the middle or at the end of paragraphs, grand activation usually has local maxima. Apparently, one of the cognitive functions of a paragraph is a threshold of activation update.
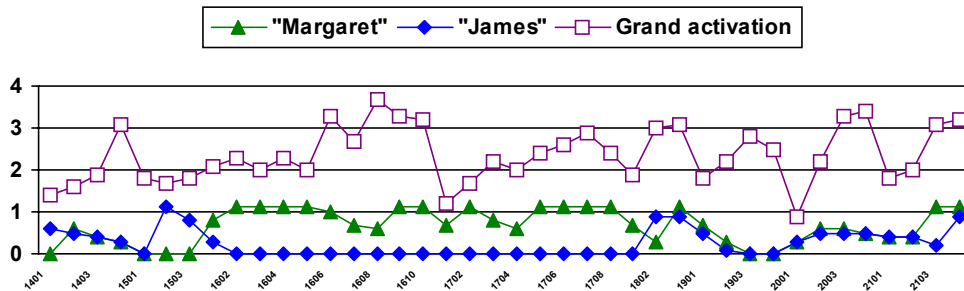

Figure 4: The dynamics of two protagonist referents' activation and of grand activation in an excerpt of English narrative (given in the Appendix)

The question of *control* of WM is the question of how information comes into WM. The current cognitive literature connects attention and WM (see e.g. (Miyake & Shah, 1999)). The issue of this connection is still debated, but the following claim seems compatible with most approaches:

➢   the mechanism controlling WM is what has long been known as attention

This claim is compatible with the already classical approaches of Baddeley (1990) and Cowan (1995), with the neurologically oriented research of Posner and Raichle (1994), and cutting-edge studies such as (McElree, 2001). According to Posner and Raichle (1994: 173), information flows from executive attention, based in the brain area known as anterior cingulate, into WM, based in the lateral frontal areas of the brain.

At the same time, as it has been convincingly demonstrated in the experimental study by Tomlin (1995), attention has a linguistic manifestation, namely grammatical roles. In many languages, including English, focally attended referents are consistently coded by speakers as the subjects of their clauses. As it has been demonstrated in the present paper, subjecthood and reduced forms of reference are causally related: antecedent subjecthood is among the most powerful factors leading to the selection of a reduced form of

reference. In both English and Russian, antecedent subjecthood can add up to 0.4 to the overall activation of a referent. In both English and Russian sample discourses, 86% of pronouns allowing no referential alternative have subjects as their antecedent.

Considered together, these facts from cognitive psychology and linguistics lead one to a remarkably coherent picture of the interplay between attention and WM, both at the linguistic and at the cognitive level. Attention feeds WM, i.e. what is attended at moment t_n becomes activated in WM at moment t_{n+1}. Linguistic moments are discourse units. Focally attended referents are typically coded by subjects; at the next moment, they become activated (even if they were not before) and are coded by reduced NPs. The relationships between attention and WM, and between their linguistic manifestations, are presented in Table 6.[10]

Moments of time (discourse units)	t_n	t_{n+1}
Cognitive phenomenon	focal attention	high activation
Linguistic reflection	mention in the subject position	reduced NP reference
Examples	Margaret, she	*she, her*

Table 6: Attention and working memory in cognition and in discourse

2.6. Conclusions about the cognitive calculative approach

The approach outlined above aims at predicting and explaining all referential occurrences in the sample discourse. This is done through a rigorous calculative methodology aiming at maximally possible predictive power. For each referent at any point in discourse, the numeric weights of all involved activation factors are available. On the basis of these weights, the integral current AS of the referent can be calculated, and mapped onto an appropriate referential device in accordance with referential strategies. The objective fluidity of the process of referential choice is addressed through the distinction between the categorical and potentially alternating referential devices. This approach allows to overcome the traditional stumbling blocks of the studies of reference: circularity and multiplicity of involved factors. The linguistic study of referential choice in discourse was based on cognitive-psychological research,

---

[10] As it has been suggested by an anonymous reviewer, this account may resemble the claims of the Centering theory on dynamics of forward- and backward-looking centers. However, we would point out that the concept of "backward-looking center" is quite different from our idea of referent activation in the subsequent discourse unit. Centering theorists posit a single backward-looking center and claim that it is the referent that discourse unit is about (see e.g. (Walker & Prince, 1996: 294-5)). Therefore, backward-looking center must be more like topic or attention focus rather than activated referent. We do not know how such concept of backward-looking center could be incorporated in the cognitively inspired model of attention-memory interplay we propose.

and it proved, in turn, relevant for the study of cognitive phenomena in a more general perspective.

3   The neural network approach

3.1   Shortcomings of the calculative approach

There are some problems with the cognitive calculative approach, especially with its calculative, or quantitative, component that was mathematically quite unversed.

First, the list of relevant activation factors may not be exactly necessary and sufficient. Those factors were included in the list that showed a strong correlation with referential choice. However, only all factors in conjunction determine the activation score, and therefore the strength of correlation of individual factors may be misleading, and the contribution of individual factors is not so easy to identify. We would like to construct an "optimal" list of factors, i.e. a model that provides maximal descriptive power (all relevant factors identified and included) and at the same time has a minimal descriptional size (just the relevant factors contained and no others).

Second, numeric weights of individual factors' values were chosen by hand which not only was a laborious task, but also did not allow judging the quality or uniqueness of the set of calculated weights.

Third, the interaction between factors was mainly additive, ignoring possible non-linear interdependencies between the factors. Non-linear dependencies are particularly probable, given that some factors interact with others (cf. the discussion of the factor of syntactic role of the linear antecedent in Section 2.4 above, whose contribution to AS depends on the linear distance).[11] Other factors might be correlated, e.g. animacy and the syntactic role of subject (the distribution of animacy and subjecthood of the antecedent vis-à-vis full NPs vs. pronouns is very similar, indicating a possible intrinsic interrelationship between these).[12] Also, from the cognitive point of view, it is unlikely that such a simple procedure as addition can adequately describe processing of activation in the brain: the basic building blocks of the brain, the nerve cells or neurons, exhibit non-linear behaviour, for example due to saturation effects. It is well

---

[11] Indeed, the attribution of different weights to the syntactic role of the linear antecedent *depending* on the linear distance in the calculative approach can already be viewed as an element of non-linear interdependencies.
[12] As a mathematical consequence, the weights attributed to animacy and antecedent subjecthood are not "stable": The model would perform almost as well if the numeric weights for these two factors were interchanged or even modified so that their sum remained the same. Thus, the concrete single weights of correlated factors have no objective importance on their own, and it is important to single out correlated factors and describe their relationship in order to ascribe an objective meaning to a combination (most simply, the sum) of their weights.

known that purely linear learning schemes cannot even solve the simple exclusive-or problem, see e.g. (Ellis & Humphreys, 1999: Ch. 2.4). For an in-depth discussion of the usefulness of non-linearity in cognitive and developmental psychology, we refer to (Elman *et al.*, 1996).

Fourth, because of the additive character of factor interaction it was very hard to limit possible activation to a certain range. It would be intuitively natural to posit that minimal activation varies between zero and some maximum, which can, without loss of generality, be assumed to be one. However, because of penalizing factors such as paragraph distance that deduct activation, it often happens that activation score turns out negative (a consequence of the simple summing in the calculative approach), which makes cognitive interpretation difficult.

In order to solve these problems, the idea to develop a more sophisticated mathematical apparatus emerged, such that:

- identification of significant factors, numeric weights, and factor interaction would all be interconnected and would be a part of the same task
- the modeling of factors would be done computationally, by building an optimal model of factors and their interaction.

There are many well-known approaches that lend themselves naturally to the problems mentioned above (e.g. variants of decisions tree algorithms, multiple non-linear regression). Since what we have in mind is to develop a quantitative cognitive model of referential choice as a long-term goal, artificial neural network models had a strong appeal to us due to their inherent cognitive interpretation (Ellis & Humphreys, 1999), even though we cannot expect a concrete cognitive model or interpretation to derive from this pilot study based on just a small data set.[13] We note that the – at first sight – less transparent representation of knowledge in a neural network, as compared to classical statistical methods, is balanced by the fact that the type of regularities it can detect in the data is less constrained.

We would like to emphasize that the primary aim of this pilot study on a quite small data set is to evaluate whether neural networks are applicable to the problem of referential choice, and if so, to lay the ground for a larger-scale study. In order to keep the present study comparable to the calculative approach, we had to use the original data set and neglected from the outset factors that already had been judged secondary.

We dispense with a more sophisticated statistical analysis of the following computer simulations since – from the point of view of rigorous statistics – the

---

[13] With respect to the small data set, we would not be better off with any other of the above mentioned methods as all of them are quite data-intensive.

data set is too small to lead to reliable results. Our intention is to get a first taste of where neural networks might take us in the analysis of referential choice.

3.2   Proposed solution: a neural network approach

In the neural network approach, we lift the requirement of complete predictiveness: we posit that referential choice can predict/explain referential choice with a degree of certainty that can be less than 100%.[14] Also, at this time, the neural network approach does not make specific claims about cognitive adequacy and activation and there is no such thing as summary activation score in this approach at its present stage. Activation factors themselves are reinterpreted as mere parameters or variables in the data that are mapped onto referential choice. We expect that at a later stage – i.e. trained on bigger data sets – the neural network approach can embrace the quantitative cognitive component.

The term *artificial neural network* or *net* denotes a variety of different function approximators that are neuro-biologically inspired (Mitchell, 1997). Their common property is that they can, in a supervised or unsupervised way, learn to classify data. For this pilot study, we decided to employ a simple feed-forward network with the back-propagation learning algorithm.

A feed-forward network consists of *nodes* that are connected by *weights*. Every node integrates the activation it gets from its predecessor nodes in a non-linear way and sends it to its successors. The nodes are ordered in layers. Numeric data is presented to the nodes in the *input layer*, from where the activation is injected into one or more hidden layers, where the actual computation is done. From there activation spreads to the *output layer*, where the result of the computation is read off. This computed output can be compared to the expected target output, and subsequently the weights are adapted so <u>as to</u> minimize the difference between actual output and target (a so-called gradient descent algorithm, of which the backpropagation algorithm is an example, for details we refer to (Ellis & Humphreys, 1999)).

In this supervised learning task, the network must learn to predict from ten factors (Table 7), whether the given referent will be realized as a pronoun or a full noun phrase. In order to input the factors with symbolic values into the net, they have to be converted into numeric values. If the symbolic values denote some gradual property such as animacy, they are converted into one real variable with values between −1 and 1. The same holds true for binary variables. When there was no a priori obvious order in the symbolic values,[15]

---

[14] This might be a desirable feature, e.g. to account for alternating referential options.

[15] For example, the factor of syntactic role can take the values "subject", "direct object", "indirect object", "possessive", etc. One might speculate that a hierarchy of these values, similar to the hierarchy of NP

they were coded unary (e.g. Syntactic Role), i.e. to every value of that factor corresponds one input node, which is set to one if the factor assumes this value and to zero otherwise.

Thus, 24 input nodes and 1 output node are needed. The output node is trained to predict whether the referent in question is realized as a full noun phrase (numeric output below 0.4) or as a pronoun (numeric output above 0.6).[16] At this point, all numeric input values were normalized to have zero mean and unit variance. This normalization ensures that all data are a priori treated on equal footing and the impact of a factor can be directly read off from the strength of the weights connecting its input node to the hidden or output layer.

Factor	Values	Coding	Input Nodes
Syntactic role	S, DO, IOag, Obl, Poss	Unary	1–5
Animacy	Human, animal, inanimate	Human: 1, animal: 0, inanimate: −1	6
Protagonisthood	Yes / no	Binary	7
Syntactic role of rhetorical antecedent*	S, DO, IOag, Obl, Poss, Pred	Unary	8–13
Type of rhetorical antecedent	Pro, FNP	Binary	14
Syntactic role of linear antecedent	S, Poss, Obl, Pred, DO, IOag	Unary	15–20
Type of linear antecedent	Pro, FNP	Binary	21
Linear distance to antecedent	Integer	Integer	22
Rhetorical distance to antecedent	Integer	Integer	23
Paragraph distance to antecedent	Integer	Integer	24

S, DO, IOag, Obl, Poss mean subject, direct object, agentive indirect object, oblique, and possessor. Pred means predicative use, Pro pronoun and FNP full noun phrase.

Table 7. Factors used in Simulation 1, their possible values and the corresponding input nodes.

3.3 Simulation 1 – full data set
A network with 24 nodes in a single hidden layer was trained on the data set of 102 items[17] from (Kibrik, 1999) (see Section 2.4) for 1000 epochs.[18] As parts of

---

[16] An output value between 0.4 and 0.6 is considered unclassified. However, this did not happen in the simulations presented here. Of course, the target values are 0 and 1 for pronouns and full NPs, respectively. Yet, for technical reasons it is preferable to admit a small deviation of the output value from the target values.

accessibility (Keenan & Comrie, 1977), might operate in referential choice. But since this is not self-evident, we code such factors unarily so that the network can find its own order of the values as relevant for the task at hand.

[17] As opposed to the study in Section 2.4, here the syntactic pronouns were included. Note that due to short linear distance all of them are easily predicted correctly.

the training are stochastic, that experiment was repeated several times. In all runs the net learned to predict the data correctly except for a small number (below six) cases. Typically, the misclassifications occurred for the same items in the data set, independently of the run. A closer analysis of a well-trained net with only four misclassifications revealed that three of them were due to referential conflict (which was not among the input factors), that is, in the situation when the full noun phrase is used only because a pronoun (otherwise expected) may turn out ambiguous.

3.4   Simulation 2 – pruning

Not only did we want our net to learn the data but also to make some statements about the importance of the input factors and their interdependency. To achieve this goal, we submitted the trained net from Simulation 1 to a *pruning procedure*, which eliminates nodes and weights from the net that contribute to the computation of the result only little or not at all. In such case, a node or weight is selected and eliminated. Then the net is retrained for 100 epochs. If net performance does not drop, the elimination is confirmed; otherwise, the deleted node or weight is restored. This procedure is repeated until no further reduction in the size of the net is possible without worsening the performance.[19]

This procedure leads to smaller nets that are easier to analyse and furthermore can reduce the dimensionality of the input data. They have a lower number of weights (i.e. a lower number of free parameters: in the case analysed here, the number of weights was reduced from 649 for the full net to 26 for the pruned net). The weights of a generic example of a pruned network trained on our data are shown in Table 8. There are no weights connecting the input nodes 3, 4, 5, 6, 11, 13, 18, 19, 20, 23 (see Table 8; the meanings of the nodes can be found in Table 7). This means that not all input factors or all their values are relevant for computing the output. Also, all but two hidden nodes have been pruned. So the two remaining suffice to model the interaction between the input factors.

Some input nodes have a direct influence on the output node (27), e.g. the node indicating that the rhetorical antecedent was a possessor (node 9). Others influence the outcome only indirectly by interacting with other nodes, e.g. paragraph distance (node 24), while yet others influence the output both

---

[18] Technical details for NN experts: learning parameter is set to 0.2; no momentum; weights were jogged every epoch by maximally 0.1%; input patterns are shuffled. The simulations are run on the SNNS network simulator (http://www-ra.informatik.uni-tuebingen.de/SNNS).

[19] More precisely, first we apply the *non-contributing units* algorithm (Dow & Sietsma, 1991), and then pruning of the minimal weight.

directly and indirectly. Some nodes enter in multiple ways that seem to cancel each other, e.g. node 14 (type of rhetorical antecedent).

Target node	Source Nodes (Weights)							
25	1 (-2.4)	2 (2.1)	8 (-1.7)	12 (1.9)	14 (-1.6)	16 (-2.4)	22 (-4.7)	24 (-4.9)
26	7 (1.7)	10 (-2.0)	12 (-5.0)	14(-1.9)	15 (2.8)	16 (-1.8)	21 (-4.2)	
27	2 (-3.7)	8 (3.9)	9 (2.0)	15 (2.7)	17 (1.8)	22 (-22.0)	25 (10.9)	26 (-10.0)

Nodes 1—24 denote the input nodes, 25 and 26 are the two remaining hidden nodes and 27 is the output node. The weights connecting a source and a target node are given in parentheses after the source node.

Table 8. Weights of a typical pruned net.

Pruning again is partly a stochastic procedure, as it for example depends ultimately on the random initialisation of the network, so we repeated the experiment until we got an impression of which factors are almost invariably included. It turned out that subject and possessor roles,[20] protagonisthood, subjecthood of the antecedent and type of antecedent are most important, and those nodes related to the rhetorical antecedent are more involved than those for the linear one. Accordingly, the most important distance is rhetorical distance. As expected, this list of factors and values coincides to a great extent with what was discovered through the trial-and-error procedure in the calculative approach. Therefore, at least qualitatively, the neural network approach is on the right track, and we can use the results of the pruning case study as a hint on how to reduce the dimensionality of the input data. This leads us to the next simulation.

3.5   Simulation 3 – reduced data set

In a third case study, we trained a similar net with 12 hidden nodes on a reduced set of only five input factors (corresponding to six input nodes): We included the values "subject" and "possessor" for syntactic role (nodes 1, 2), protagonisthood (node 3), both when the rhetorical antecedent was a subject (node 4), and when it was realized as a pronoun or full NP (node 5), and rhetorical distance (node 6). The new net had 12 hidden nodes, corresponding to 103 weights. On this reduced net, we executed the back-propagation learning algorithm for 500 epochs and then pruning (50 epochs retraining for each pruning step) with the same parameters as before. We ended up with a small net (23 parameters), shown in Figure 5, that classified only 8 out of 102 items

---

[20] Interestingly, some hints on the difference in the usage of argumental and possessive pronouns were observed already during the original work on the calculative approach. The fact that the networks themselves frequently keep the input for the possessive role can be viewed as a corroboration of this thought, and also as a proof that neural networks can be used as an independent tool for discovering regularities in the data. Work focusing on this differentiation is underway.

wrongly. Note that all remaining factors interact strongly, except for protagonisthood (node 3), which has been pruned away.
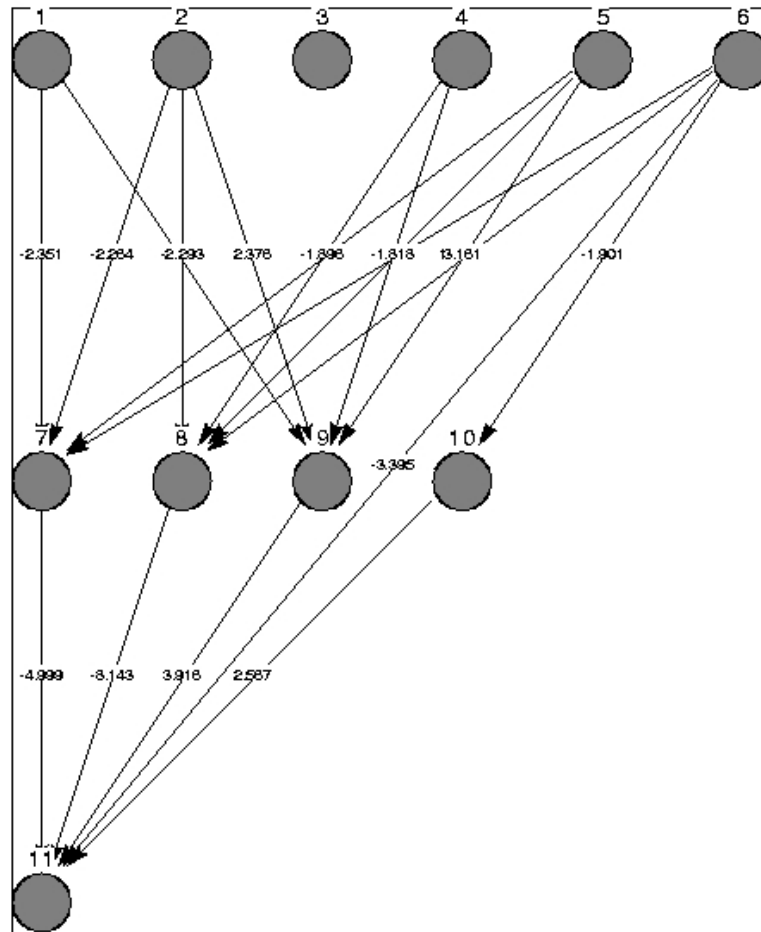
The circles denote the nodes, the arrows the weights connecting the nodes, to which the weight strength is added as a real number. Nodes 1–6 are input nodes, 7–10 the nodes in the hidden layer, and node 11 is the output.

Figure 5. Net from Simulation 3

3.6   Simulation 4 – cheap data set

Reliable automatic annotators for rhetorical distance and consequently for all factors related to the rhetorical antecedent, as well as for protagonisthood, are not available. Since these factors require comprehension of the contents of the

text, they must be annotated by human experts and are therefore costly. So we decided to replace the rhetorical factors included in Simulation 3 by the corresponding linear ones and protagonisthood by animacy. Keeping the six input nodes as before, we added a seventh one to indicate that the linear antecedent was a possessor and an eighth one for paragraph distance to help the net to overcome the smaller amount of information that is contained in the linear antecedent factors. Training and pruning proceeded as before.

One typical resulting network in this case had 32 degrees of freedom. Again, animacy, which had been substituted for protagonisthood, is disconnected from the rest of the net. On the 102 data items, the net produced only six errors (three are due to referential conflict).

Thus, even though the logical structure of the factors and their values was considerably simplified, and none of the factors included that relate to the rhetorical antecedent, the accuracy (six errors versus four with the full set of factors) did not deteriorate dramatically.

3.7   Comparison to the calculative approach

In the calculative model discussed in Section 2.4 above, referential choice was modelled by 11 factors using 32 free parameters (counting the number of the different numeric weights for all factors and their values). The activation score allowed a prediction of the referential choice in five categories. In our study with neural networks, we modelled only a binary decision (full NP/pronoun) and lifted the requirement of cognitive adequacy. The smallest net in the study, in simulation 3, had only 23 free parameters (weights), 5 input factors, and the best net on the full set of input factors, in Simulations 1 and 2, misclassified only four items, having 26 free parameters.

Even though the accuracy dropped in the neural network approach (using a reduced set of input factors) as compared to the calculative approach (with the full set of input factors), the descriptional length (measured in the number of free model parameters) was reduced by approximately one third and thus yields in this sense a more compact description of the data.

These findings are important in the following respects. First, we can find a smaller set of factors that still allows a relatively good prediction of referential choice, but it is much less laborious to extract from a given corpus, thus making the intended large-scale study feasible. Second, we can reduce the descriptional length without a too severe drop in accuracy. This means that the networks were able to extract the essential aspects of referential choice since about 100 instances can be described by only 23 parameters. Compare this to the worst case in which a learning algorithm needs about 100 free parameters to describe 100 instances. In such case, the algorithm would not have learnt

anything essential about referential choice, because it would be merely the list of the 100 instances. The ratio of the number of parameters to the size of data set has a long tradition of being used for judging a model's quality. A high value of this ratio is an indicator for overfitting[21] (see any standard textbook on statistics).

In large-scale studies, which are due to follow this pilot study, we expect to construct models with an even better ratio of descriptional length to the size of data set.

3.8   Comparison to (Strube and Wolters, 2000)

As it has been pointed out above, there are relatively few studies of referential choice – most authors are interested in resolution of anaphoric devices. Furthermore, there are almost no studies that would attempt to integrate multiple factors affecting reference. However, we are familiar with one study that is remarkably close in spirit to ours, namely (Strube & Wolters, 2000). Strube and Wolters use a similar list of factors as the calculative approach discussed above, except that the costly factors related to the rhetorical antecedent are missing. They analyse a large corpus with several thousand of referring expressions for the categorical decision (full NP/pronoun) using logistic regression. The logistic regression is a form of linear regression adapted for a binary decision.

Factor interaction and non-linear relations are thus not accounted for in their model, and they present no cognitive interpretation of their model either. Still, the gist and intention of their and our studies – independently developed – largely agree, which provides evidence for the usefulness and appropriateness of quantitative approaches towards referential choice.

4 Conclusion and outlook

In Section 3, we reported a pilot study testing whether artificial neural networks are suitable to process our data. We trained feed-forward networks on a small set of data. The results show that the nets are able to classify the data almost correctly with respect to the choice of referential device. A pruning procedure enabled us to single out five factors that still allowed for a relatively good prediction of referential choice. Furthermore, we demonstrated that costly input factors such as rhetorical distance to the antecedent could be replaced by those related to the linear antecedent, which can be more easily collected from a large corpus.

---

[21] Overfitting means sticking too closely to the peculiarities of a given training set and not finding the underlying general regularities. Overfitting is roughly the opposite of good generalization of unknown data.

Because of the small amount of data for this pilot study, the result must be taken with due care. But these results encourage us to further develop this approach.

Future work will include a study of a larger data set. This is necessary since neural networks as well as classical statistics need a large amount of data to produce reliable results that are free of artefacts. In our corpus, some situations (i.e. an antecedent that is an indirect object) appear only once, so that no generalization can be made. In a larger study, the advantages of the neural network approach can be used fully.

We also aim at reintroducing a cognitive interpretation at a later stage, and we want to work with different network methods, that not only allow dimensional reduction and data learning, but also an easy way to explicitly extract the knowledge from the net in terms of more transparent symbolic rules (see e.g. (Kolen & Kremer, 2001)).

Furthermore, we feel the need not only to model a binary decision (full NP/pronoun), but also to have a more fine-grained analysis. The calculative approach of Section 2.4 has done the first steps in this direction, allowing for five different categories that not only state that a pronoun or a full NP is expected, but also to what degree a full NP in a particular situation can be replaced by a pronoun and vice versa.

A statistical interpretation of referential choice can be suggested if a human expert judges that a particular full NP could be replaced by a pronoun, s/he must have experienced that in a very similar situation where the writer did indeed realize the other alternative. The expert will be more certain that substitution is suitable if s/he has often experienced the alternative situation. Thus we think it is promising to replace the five categories discussed in Section 2.4 by a continuous resulting variable that ranges from 0 to 1 and is interpreted as the probability that referential choice realizes a pronoun in the actual situation: 1 means a pronoun with certainty, 0 means a full NP with certainty, and 0.7 means that in 70% instances a pronoun is realized and a full NP in the remaining 30% instances.

As an anonymous reviewer pointed out to us, there is an interesting potential application of neural network-based models of referential choice to anaphor resolution. Consider a knowledge-poor anaphor resolution algorithm as a quick-and-dirty first pass that suggests several potential referents for a pronominal mention. Counterchecking the referent mentions in a second pass, a suggested referent could be ruled out if the network does not predict a pronominal mention for it at the point in question. The advantage over anaphor resolution algorithms based purely on classical methods would be that

computations in a neural network are really fast compared to algorithmic and symbolic computing once the training of the network is finished.

Acknowledgements

References

Anderson, J. R. 1990. *Cognitive Psychology and its Implications*. 3rd ed. New York: W. H. Freeman & Co.

Baddeley, A. 1986. *Working Memory*. Oxford: Clarendon Press.

---------- 1990. *Human Memory: Theory and Practice*. Needham Heights: Allyn and Bacon.

Botley, S. and A. M. McEnery (eds.) 2000. *Corpus-based and Computational Approaches to Discourse Anaphora*. Amsterdam: John Benjamins.

Carlson, L., D. Marcu and M. E. Okurowski 2003. "Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory." Jan van Kuppevelt and Ronnie Smith (eds.) *Current Directions in Discourse and Dialogue*. Dordrecht: Kluwer. To appear.

Chafe, W. 1994. *Discourse, Consciousness, and Time. The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: University of Chicago Press.

Clifton, C. Jr. and F. Ferreira 1987. "Discourse Structure and Anaphora: Some Experimental Results." M. Coltheart (ed.) *Attention and Performance XII*. Hove: Erlbaum.

Cornish, F. 1999. *Anaphora, Discourse, and Understanding. Evidence from English and French*. Cambridge: Cambridge University Press.

Cowan, N. 1995. *Attention and Memory: An Integrated Framework*. New York: Oxford: Oxford University Press.

---------- 2000. "Childhood Development of some Basic Parameters of Working Memory." Schroeger, Mecklinger, and Friederici 2000. 139–160.

Dale, R. 1992. *Generating Referring Expressions*. Cambridge: MIT Press

Dow, R. and J. Sietsma. 1992. "Creating Artificial Neural Networks that Generalize." *Neural Networks* 4(1): 67–79.

Ellis, R. and G. W. Humphreys 1999. *Connectionist Psychology*. East Sussex: Psychology Press.

Elman, J. L., E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi and K. Plunkett 1996. *Rethinking Innateness – A Connectionist Perspective on Development*. Cambridge: MIT Press.

Fox, B. 1987a. *Discourse Structure and Anaphora*. Cambridge: Cambridge University Press.

---------- 1987b. "Anaphora in Popular Written English Narratives." R. Tomlin (ed.) *Coherence and Grounding in Discourse*. Amsterdam: John Benjamins. 157-174.

Gathercole, S. E. (ed.) 1996. *Models of Short-term Memory*. East Sussex: Psychology Press.

Gernsbacher, M. A. 1990. *Language Comprehension as Structure Building*. Hillsdale: Erlbaum.

Givón, T. (ed.) 1983. *Topic Continuity in Discourse. A Quantified Cross-language Study*. Amsterdam: John Benjamins.

---------- 1990. *Syntax: A Functional-typological Introduction*. Vol. 2. Amsterdam: John Benjamins.

---------- 1995. *Functionalism and Grammar*. Amsterdam: John Benjamins.

Gordon, P. C., B. J. Grosz and L. A. Gilliom 1993. "Pronouns, Names, and the Centering of Attention in Discourse." *Cognitive Science* 17: 311–347.

Grimes, J. 1978. *Papers in Discourse*. Arlington: SIL.

Gundel, J. K., N. Hedberg and R. Zacharski 1993. "Cognitive Status and the Form of Referring Expressions in Discourse." *Language* 69: 274–307.

Keenan, E. and B. Comrie 1977. "Noun Phrase Accessibility and Universal Grammar." *Linguistic Inquiry* 8-1: 63–98.

Kibrik, A. A. 1991. "Maintenance of Reference in Sentence and Discourse." Winfred P. Lehmann and Helen J. Hewitt (eds.) *Language Typology*. Amsterdam: John Benjamins. 57-84.

---------- 1996. "Anaphora in Russian Narrative Prose: A Cognitive Account." Barbara Fox (ed.) *Studies in Anaphora*. Amsterdam: John Benjamins.255-304.

---------- 1999. "Reference and Working Memory: Cognitive Inferences from Discourse Observation." K. van Hoek, A. A. Kibrik and L. Noordman (eds.) *Discourse Studies in Cognitive Linguistics*. Amsterdam: John Benjamins.29-52.

---------- 2000. "A Cognitive Calculative Approach Towards Discourse Anaphora." P. Baker, A. Hardie, T. McEnery and A. Siewierska (eds.) *Proc. of the Discourse Anaphora and Anaphor Resolution Conference (DAARC2000)*. Lancaster: University Centre for Computer Corpus Research on Language.72-82.

Kolen, J. F. and S. C. Kremer (eds.) 2001. *A Field Guide to Dynamical Recurrent Networks*. New York: IEEE Press.

Mann, W., C. Matthiessen and S. A. Thompson 1992. "Rhetorical Structure Theory and Text Analysis." W. Mann and S. Thompson (eds.) *Discourse Description. Diverse Linguistic Analyses of a Fund-raising Text*. Amsterdam: John Benjamins.39-78.

Marslen-Wilson, W., E. Levy and L. K. Tyler 1982. "Producing interpretable discourse: The Establishment and Maintenance of Reference." *Speech, Place, and Action. Studies in Deixis and Related Topics*, ed. R.J. Jarvella and W. Klein, 339-378. Chichester: Wiley.

McElree, B. 2001. "Working Memory and Focal Attention." *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Vol. 27, No. 3, 817-835.

Mitchell, T. M. 1997. *Machine Learning*. Boston: McGraw-Hill.

Miyake, A. and P. Shah (eds.) 1999. *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge: Cambridge University Press.

Posner, M. I. and M. E. Raichle. 1994. *Images of Mind*. New York: Scientific American Library.

Schroeger, E., A. Mecklinger and A. D. Friederici (eds.) 2000. *Working on Working Memory*. Leipzig: Leipziger Universitätsverlag.

Shah, P. and A. Miyake 1999. "Models of Working Memory: An Introduction." Miyake and Shah (eds.). 1–26.

Smith, E. E. and J. Jonides. 1997. Working Memory: A View from Neuroimaging. *Cognitive Psychology* 33: 5–42.

Strube, M. and M. Wolters 2000. "A Probabilistic Genre-Independent Model of Pronominalization." *Proc. of the 1st Meeting of the North American Chapter of the ACL. Seattle, WA, USA, April 29 - May 4, 2000,* 18–25.

Tomlin, R. 1987. "Linguistic Reflections of Cognitive Events." R. Tomlin (ed.) *Coherence and Grounding in Discourse*. Amsterdam: John Benjamins. 455-479.

---------- 1995. "Focal Attention, Voice and Word Order: An Experimental Cross-linguistic Study." P. Downing and M. Noonan (eds.) *Word Order in Discourse*. Amsterdam: John Benjamins. 517–554.

---------- and M.-M. Pu 1991. "The Management of Reference in Mandarin Discourse." *Cognitive Linguistics* 2: 65–93.

Toole, J. 1996. "The Effect of Genre on Referential Choice." T. Fretheim and J. Gundel (eds.) *Reference and Referent Accessibility*. Amsterdam: John Benjamins. 263–290.

van Hoek, K. 1997. *Anaphora and Conceptual Structure*. Chicago: University of Chicago Press.

Van Valin, R. D. Jr. 1993. "A Synopsis of Role and Reference Grammar." R. D. Van Valin, Jr. (ed.) *Advances in Role and Reference Grammar*. Amsterdam: John Benjamins.

Velichkovsky, B. M., B. H. Challis and M. Pomplun 1995. "Arbeitsgedächtnis und Arbeit mit dem Gedächtnis: Visuell-räumliche und weitere Komponenten der Verarbeitung." *Zeitschrift für experimentelle Psychologie* 42 (4): 672-701.

Vonk, W., L. G. M. M. Hustinx and W. H. G Simons 1992. "The Use of Referential Expressions in Structuring Discourse." *Language and Cognitive Processes* 7: 301–333.

Walker, M. A. and E. F. Prince 1996. "A Bilateral Approach to Givenness: A Hearer-status Algorithm and a Centering Algorithm." T. Fretheim and J. Gundel (eds.) *Reference and Referent Accessibility*. Amsterdam: John Benjamins. 291–306.

Appendix: An Excerpt from an English Narrative
("The Maggie B." by Irene Haas)

1401	Margaret and James were cold.
1402	The sky grew darker.
1403	The goat and chickens fled into their little shelter,
1404	the toucan flew screeching into the cabin.
1501	James started to cry.
1502	A storm was coming!
1503	Margaret must make the boat ready at once.
1601	She took in the sail
1602	and tied it tight.
1603	She dropped the anchor
1604	and stowed all the gear,
1605	while rain drummed on the deck
1606	and thunder rumbled above her.
1607	Lightning split the sky
1608	as she ran into the cabin
1609	and slammed the door against the wet wind.
1610	Now everything was safe and secure.
1701	When she lit the lamps,
1702	the cabin was bright and warm.
1703	It was nearly suppertime
1704	so Margaret mixed up a batch of muffins
1705	and slid them into the oven.
1706	She sliced some peaches

1707   and put cinnamon and honey on top,
1708   and they went into the oven, too.
1801   James was given a splashy bath in the sink.
1802   Margaret dried him in a big, warm towel,
1803   and then supper was ready.
1901   Outside, the wind howled like a pack of hungry wolves.
1902   Rain lashed the windowpanes.
1903   But the sturdy little Maggie B. kept her balance
1904   and only rocked the nicest little bit.
2001   Margaret and James ate the beautiful sea stew
2002   and dunked their muffins in the broth,
2003   which tasted of all the good things that had cooked in it.
2004   For dessert they had the peaches with cinnamon and honey, and glasses of warm goat's milk.
2101   When supper was over,
2102   Margaret played old tunes on her fiddle.
2103   Then she rocked James in his cradle
2104   and sang him his favorite song.